American Journal of Applied Sciences 9 (2): 259-264, 2012
ISSN 1546-9239

A Control of Fundamental Frequency Contour for Hidden Markov Model-Based Thai Speech Synthesis

Suphattharachai Chomphan
Department of Electrical Engineering,
Faculty of Engineering at Si Racha, Kasetsart University,
199 M.6, Tungsukhla, Si Racha, Chonburi, 20230, Thailand

Abstract: Problem statement: In the conventional HMM-based speech synthesis system for Thai, there is no control of fundamental frequency control in the synthesis stage. The tone correctness of the synthesized speech is unacceptable due to the imbalance of training data of all tones. **Approach:** This study proposes a mathematical model to control the F_0 contour of the synthesized speech. This control is proposed to correct only some distorted segments of the F_0 contour which occur within some syllables due to lacking of training data for some tones. **Results:** An experimental result compares F_0 contours between those of synthesized speech with and without tone-type questions; furthermore the size of Thai speech corpus is varied to investigate the synthesized speech quality. A mathematical model is applied to control the F_0 contour. By using the proposed control, the correction of the F_0 contour is obviously shown in the experimental results. **Conclusion:** The control of F_0 contour has been proposed. It can noticeably improve the tone correctness of the synthesized speech.

Key words: Frequency contour, thai speech, speech synthesis, HMM-based speech synthesis, tone correctness, Text-To-Speech (TTS), National Electronics and Computers Technology Center (NECTEC), F_0 contour

INTRODUCTION

Speech synthesis is an important technology for realizing natural human-computer interaction. For this purpose, Text-To-Speech (TTS) systems are required to have an ability to generate speech with arbitrary speaker’s voice characteristics and various speaking styles. A number of TTS techniques have been proposed and state-of-the-art TTS systems based on unit selection and concatenation can generate natural sounding speech. However, it is a difficult problem to synthesize speech with various voice characteristics and various speaking styles.

HMM-based TTS system in which each speech synthesis unit is modeled by HMM is proposed in the recent years (Masuko *et al*., 1996; Yoshimura *et al*., 1999). A distinctive feature of the system is that the speech parameters used in the synthesis stage are generated directly from HMMs by using a parameter generation algorithm (Tokuda *et al*., 1999). Since the HMM-based TTS system uses HMMs as the speech units in both modeling and synthesis, the voice characteristics of synthetic speech can be changed by transforming HMM parameters appropriately.

In the development of Thai speech synthesis, a TTS synthesis system based on unit selection is initially implemented by (Chomphan and Kobayashi, 2007a). Subsequently, a TTS synthesis system based on unit selection with TD-PSOLA technique is developed by National Electronics and Computers Technology Center (NECTEC) in 2003 (Hansakunbuntheung *et al*., 2005). Since Thai is a tonal language, this study is proposed to implement Thai speech synthesis based on HMM which has the ability of synthesizing speech with various voice characteristics and various speaking styles, moreover an additional control of fundamental frequency contour of the synthetic speech is proposed. The proposed controlling is done based on structural modeling of voice fundamental frequency contours which is previously achieved with Mandarin speech (Ni and Hirose, 2006).

MATERIALS AND METHODS

HMM-Based speech synthesis: A block-diagram of the HMM-based TTS system is shown in Fig. 1. The system consists of two stages including the training stage and the synthesis stage (Yamagishi *et al*., 2003).

In the training stage, mel-cepstral coefficients are extracted at each analysis frame as the static features from the speech database. Then the dynamic features, i.e., delta and delta-delta parameters, are calculated from the static features. Spectral parameters and pitch observations are combined into one observation vector frame-by-frame and speaker dependent phoneme

HMMs are trained using the observation vectors. To model variations of spectrum, pitch and duration, phonetic and linguistic contextual factors, such as phoneme identity factors, are taken into account (Yoshimura *et al*., 1999). Spectrum and pitch are modeled by multi-stream HMMs and output distributions for spectral and pitch parts are continuous probability distribution and Multi-Space Probability Distribution (MSD) (Tokuda *et al*., 1999), respectively. Then, a decision tree based context clustering technique is separately applied to the spectral and pitch parts of context dependent phoneme HMMs (Yoshimura *et al*., 1998). Finally state durations are modeled by multi-dimensional Gaussian distributions and the state clustering technique is also applied to the duration distributions (Yoshimura *et al*., 1998).

In the synthesis stage, first, an arbitrary given text to be synthesized is transformed into context dependent phoneme label sequence. According to the label sequence, a sentence HMM, which represents the whole text to be synthesized, is constructed by concatenating adapted phoneme HMMs. From the sentence HMM, phoneme durations are determined based on state duration distributions (Yoshimura *et al*., 1998). Then spectral and pitch parameter sequences are generated using the algorithm for speech parameter generation from HMMs with dynamic features (Tokuda *et al*., 1999). Finally by using MLSA filter (Chomphan and Kobayashi, 2007b) speech is synthesized from the generated mel-cepstral and pitch parameter sequences.

Thai Speech Attributes: In tonal languages, such as Thai, a syllable is composed of consonants, vowels and tone (Chomphan, 2010). The basic Thai textual syllables can be shown in Fig. 2, where Ci, V, Cf and T denotes an initial consonant, a vowel, a final consonant and a tone respectively. The significant difference between tonal and toneless language is the syllable tone, where meaning of a syllable changes as the syllable tone changes (Thathong *et al*., 2000; Chomphan, 2010). Table 1 summarizes the number of the Thai characters and phones according to each part of syllables.

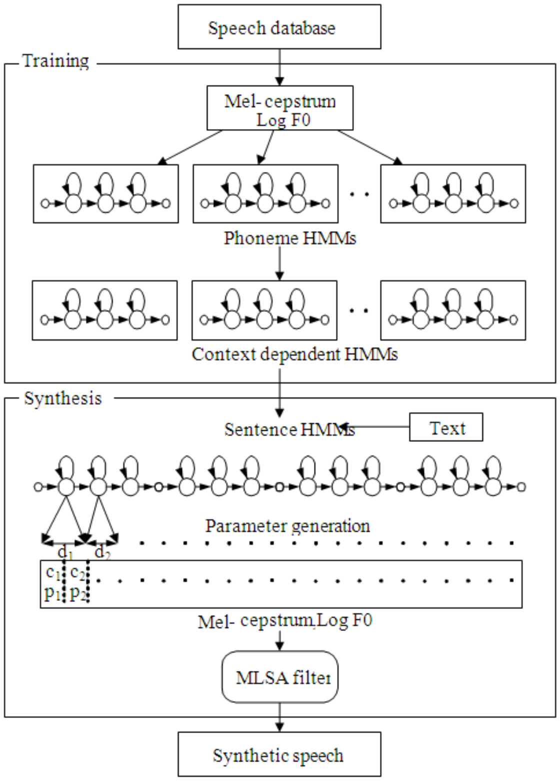

Fig. 1: A block diagram of an HMM-based speech synthesis system

T
Ci(G)V(V)Cf

Fig. 2: Thai syllable structure

Table 1: The number of Thai characters and phones

Type	Character	Phone
Initial consonant (Ci)	44	38
Vowel (V)	16	24
Final consonant (Cf)	37	9
Tone (T)	4	5

In Thai language, there are tones including middle tone, low tone, falling tone, high tone and rising tone. Each syllable tone can be characterized by its corresponding fundamental frequency contour which is depicted in Fig. 3. Each contour line is constructed by plotting the voice fundamental frequency extracted periodically via the normalized syllable duration.

F_0 contour or tone controlling by structural modeling: In the training stage of HMM-based TTS system, the sufficiency of training utterances which cover all phones in Thai is important. In some systems, lacking of training utterances can be found, consequently some syllables of the synthesized speech may convey wrong tone. An example of wrong syllable tone synthesized from the system is depicted in Fig. 4. At the first syllable (ผล, /phŏn /), there are significant difference between the two lines, one conveys true characteristic of rising tone while another conveys wrong characteristic.

Therefore, this study proposed a controlling method of fundamental frequency in the synthesis stage. This controlling method is based on the structural modeling of voice fundamental frequency contours. By applying the assumption that the behavioral characteristics of vocal-fold elongation in vibration could be approximated by those of a simple forced vibrating system, the voice fundamental frequency contour can be modeled in mathematical term (Fujisaki *et al*., 1990). Let F_0 (t) represent an F_0 contour as a function of time t in vocal range of [F_{0b}, F_{0t}]. Assume λ (t) to indicate a sequence of virtual tone graphs in λ-time space to specify to underlying lexical tone structures. Additionally assume a latent scale ζ (t) to characterize the intonation components. Thus, the F_0 contour on the logarithmic scale of fundamental frequency is expressed as a scale transformation from λ $(t)_{to}$ F_0 (t), corresponding to the syllabic tones fitting themselves with sentence intonation in vocal range:

$$\frac{\ln F_0(t)-\ln f_{0_b}}{\ln f_{0_t}-\ln f_{0_b}}=\frac{A(\lambda(t),\zeta(t))-A(\lambda_b,\zeta(t))}{A(\lambda_t,\zeta(t))-A(\lambda_b,\zeta(t))},\text{fort}\geq 0 \qquad (1)$$

$$A(\lambda,\zeta)=\frac{1}{\sqrt{(1-(1-2\zeta^2)\lambda)^2+4\zeta^2(1-2\zeta^2)\lambda}},\text{for}\lambda\geq 1 \qquad (2)$$

Here, A (λζ) takes the right arm of a resonance curve (λ≥1). Equation 1 and 2 jointly indicate a structural formulation of the control process of coupling the syllabic tones and sentence intonation together to form a final sentence melody. F_0 (t) can be reformulated from Eq. 1 as Eq. 3:

$$F_0(t)=\exp\left(\frac{A(\lambda(t),\zeta(t))-A(\lambda_b,\zeta(t))}{A(\lambda_t,\zeta(t))-A(\lambda_b,\zeta(t))}\times\ln\frac{f_{0_t}}{f_{0_b}}+\ln f_{0_b}\right) \qquad (3)$$

A local control mechanism will generate local F_0movements related to the tone graphs and thus is specified by λ (t) as a function of time t. A term local F_0movement here indicates either a simple rising or falling movement; a flat is then treated as a specific rising or falling movement.

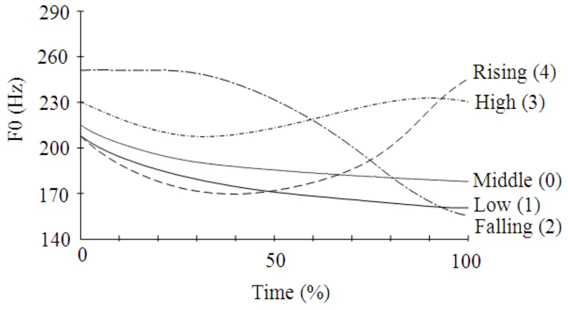

Fig. 3: Fundamental frequency contours of 5 tones in Thai

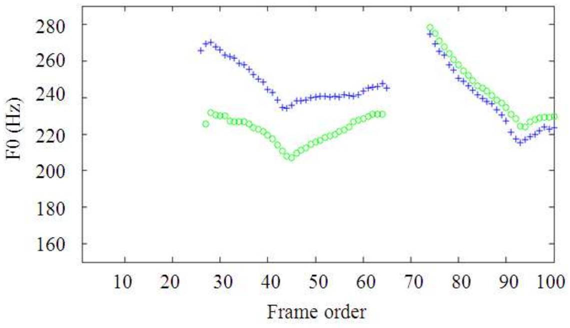

Fig. 4: Examples of fundamental frequency contours of synthesized speech (+ for wrong tone, o for correct tone)

A simple rising/falling movement is parameterized in so-called λ-time space by a passive state indicated by $\lambda_p$at t = 0the beginning time of a local F_0movement), active transition amplitude Δλand transition time Δt. Inspired by accent control mechanism of the Fujisaki model, the control mechanism of simple rising/falling movements as Eq. 4:

$$\lambda(t)=\lambda_p+\Delta\lambda\left(1-\left(1+\frac{4.8}{\Delta t}\right)e^{-\frac{4.8}{\Delta t}t}\right),\text{for } t\geq 0 \qquad (4)$$

To control the voice fundamental frequency contour of the wrong tone as depicted in Fig. 4, it can be done by adjusting the following parameters appropriately, [F_{0b}, F_{0t}] bottom and top frequencies of the vocal range in hertz, ζ (t) latent scales or intonation components:

(t_{pi},λ_{pi}) = ith peak coordinate in λ-time space; t_{p0} = 0; and t_{pn+1}=∞
$\Delta\lambda_I$ = ith peak active transition amplitude and
Δt_I = ith peak transition time

RESULTS

Experimental conditions: A set of phonetically balanced sentences of Thai speech database from National Electronics and Computer Technology Center (NECTEC) is used for training HMMs. The whole sentence text was collected from Thai part-of-speech tagged corpus, named ORCHID. The speech in the database is uttered by a professional female speaker with clear articulation and standard Thai accent. The text dependent phoneme labels are extracted based on the phoneme labels and linguistic information included in the database. There are almost 79 phonemes including silence and pause.

Speech signal were sampled at a rate of 16 kHz and windowed by a 25 m sec Blackman window with a 5 ms shift. Then mel-cepstral coefficients were extracted by mel-cepstral analysis. The feature vectors consisted of 25 mel-cepstral coefficients including the zeroth coefficient, logarithm of fundamental frequency and their delta and delta-delta coefficients (Zen *et al*., 2004; Chomphan and Kobayashi, 2007a).

We used 5-state left-to-right HMMs in which the spectral part of the state is modeled by a single diagonal Gaussian output distribution (Chomphan and Kobayashi, 2007a). The number of training utterances is varied as 500, 1000, 1500, 2000 and 2500 sentences.

Subjective evaluations of synthesized speech: First, the naturalness of the synthesized speech generated from 6 approaches; 5 are from the HMM-based system set up by varying number of training utterances and another one is from the unit selection approach with the corpus size of 5200 sentences, was evaluated by a paired comparison test. The subjects were nine Thai persons. They were presented a pair of speech synthesized from different approaches in random order and then asked which one sounded more natural. For each subject, five test sentences were chosen at random from 25 test sentences which were not contained in the training sentences. Figure 5 and 6 show the preference scores.

Effects of F_0 contour controlling: By adjusting the parameters of the modeled described earlier, while F_0b, F_0t,ζ (t) are set as constant values at 80Hz, 420Hz and 0.156 respectively, it can be seen that we can reshape the fundamental frequency contour to obtain the correct tone characteristics. Figure 7 shows an example of how the contour of fundamental frequency is reshaped appropriately.

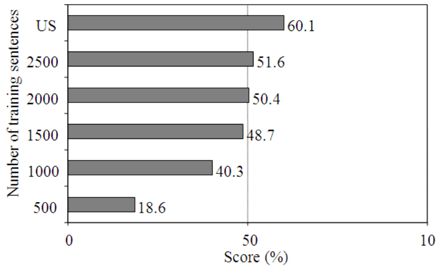

Fig. 5: Evaluation of naturalness of Thai HMM-based system and unit selection (US) approach

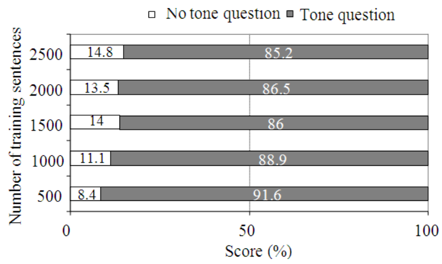

Fig. 6: Evaluation of correction of syllable tone of Thai HMM-based approach

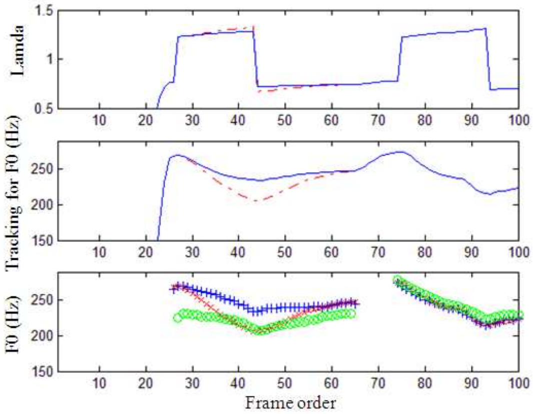

Fig. 7: Upper: λ parameter (solid line for no adjustment, dash line for adjustment), Middle: Tracking line for F_0 contours (solid line for no adjustment, dash line for adjustment), Lower: F_0 contours (+ for wrong tone, o for correct tone, x for parameter adjustment).

DISCUSSION

From Fig. 5, it can be seen that the more the number of training sentences is increased the more the naturalness of the synthesized speech is obtained. Although, the score of unit selection approach is above of HMM-based approach with 2500 training sentences, the HMM-based approach can be further developed to synthesize the speech with various voice characteristics and various speaking styles as mentioned earlier. It can be said that HMM-based approach is newly constructed for Thai language. Moreover, we expect that the system be further improved in the near future.

Secondly, the subjective evaluation of tone question in the context clustering stage was conducted. The correction of the syllable tone of the synthesized speech generated from 2 systems was evaluated by a paired comparison test. The first system has tone question in the context clustering process, meanwhile another system has no tone question.

From Fig. 6, it can be seen that the score of the system with tone question is considerably superior than that of the system without tone question for every number of training sentences. When increasing the number of training sentences, the percentage score of no tone question case increases. The reason is that the lacking problem of training syllable tones is relieved.

From Fig. 7, it is obviously seen that this approach can be embedded into the automatic control of the misshaped tone of the synthesized speech with Thai HMM-based synthesis system.

CONCLUSION

In this study, an HMM-based Thai speech synthesis is presented. Thai speech characteristic is investigated and subsequently the conventional HMM-based synthesis system is modified according the tonal attributes of Thai. It has been found that the number of training sentences affected the naturalness of the synthesized speech while the tone information affected significantly with the output synthesized speech. Moreover, a functional model is applied to control the F_0 contour. The purpose of the control is to correct the distorted segments of the F_0 contour.

ACKNOWLEDGEMENT

The researchers are grateful to Kasetsart University at Si Racha campus for the research scholarship through the board of research.

REFERENCES

Chomphan, S. and T. Kobayashi, 2007a. Implementation and evaluation of an HMM-based Thai speech synthesis system. Proceedings of the 8th Annual Conference of the International Speech Communication Association, Aug. 27-31, Antwerp, Belgium, pp: 2849-2852.

Chomphan, S. and T. Kobayashi, 2007b. Design of tree-based context clustering for an HMM-based Thai speech synthesis system. Proceedings of the 6th ISCA Workshop on Speech Synthesis, Aug. 22-24, Bonn, Germany, pp: 160-165.

Chomphan, S., 2010. Multi-pulse based code excited linear predictive speech coder with fine granularity scalability for tonal language. J. Comput. Sci., 6: 1267-1271. DOI: 10.3844/jcssp.2010.1267.1271

Fujisaki, H., K. Hirose, P. Halle and H. Lei, 1990. Analysis and modeling of tonal features in polysyllabic words and sentences of the standard Chinese. Proceedings of the 1st International Conference on Spoken Language Processing, Nov. 18-22, ICSA Archive, Kobe, Japan, pp: 841-844.

Hansakunbuntheung, C., A. Rugchatjaroen and C. Wutiwiwatchai, 2005. Space reduction of speech corpus based on quality perception for unit selection speech synthesis. National Electronics and Computer Technology Center.

Masuko, T., K. Tokuda, T. Kobayashi and S. Imai, 1996. Speech synthesis using HMMs with dynamic features. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, May 7-10, IEEE Xplore Press, Atlanta, USA., pp: 389-392. DOI: 10.1109/ICASSP.1996.541114

Ni, J. and K. Hirose, 2006. Quantitative and structural modeling of voice fundamental frequency contours of speech in Mandarin. Speech Commun., 48: 989-1008. DOI: 10.1016/j.specom.2006.01.002

Thathong, U., S. Jitapunkul, V. Ahkuputra, E. Maneenoi and B. Thampanitchawong, 2000. Classification of Thai consonant naming using Thai tone. Proceedings of the 6th International Conference on Spoken Language Processing, Oct. 16-20, ICSA Archive, Beijing, China, pp: 47-50.

Tokuda, K., T. Masuko, N. Miyazaki and T. Kobayashi, 1999. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 15-19, IEEE Xplore Press, Phoenix, USA., pp: 229-232. DOI: 10.1109/ICASSP.1999.758104

Yamagishi, J., T. Masuko, K. Tokuda and T. Kobayashi, 2003. A training method for average voice model based on shared decision tree context clustering and speaker adaptive training. Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Apr. 6-10, IEEE Xplore Press, Hong Kong, pp: 716-719. DOI: 10.1109/ICASSP.2003.1198881

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1998. Duration modeling for HMM-based speech synthesis. Proceedings of the 5th International Conference on Spoken Language Processing, Nov. 30-Dec. 4, ICSA Archive, Sydney, Australia.

Yoshimura, T., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 1999, Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis. Proceedings of 6th European Conference on Speech Communication and Technology, Sep. 5-9, ICSA Archive, Budapest, Hungary, pp: 2347-2350.

Zen, H., K. Tokuda, T. Masuko, T. Kobayashi and T. Kitamura, 2004. Hidden semi-Markov model based speech synthesis. Proceedings of the 8th International Conference on Spoken Language Processing, Oct. 4-8, ICSA Archive, Jeju Island, Korea, pp: 1393-1396.